DSC 424
IBM HR Attrition Data Analysis

Team: Data Crew

Denvir Gama, Erik Pak, Jiten Mishra, Zachary Hollis

DePaul University
Advance Data Analysis and Regression 424
Dr. John McDonald
June 06, 2023

Table of Contents

Table of Figures

Abstract

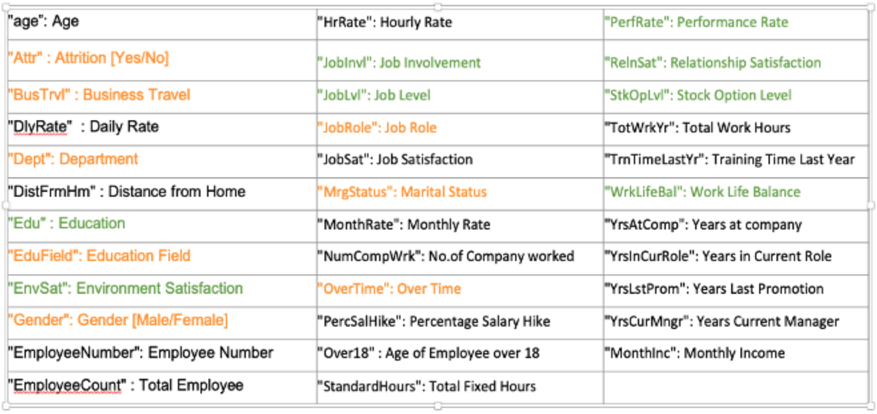
This study presents a research approach for measuring the performance of a model predicting Attrition in IBM company. The performance metrics extracted from the dataset were modeled., which captures 74% of the correctness of the attrition prediction. The model was built using different techniques listed in this document. The analysis of this model information may be used by management to make decisions to mitigate attrition risks.

Introduction

Attrition is a problem that impacts all businesses, irrespective of geography, industry, and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of and minimizing staff attrition. The goal is to analyze the predictive modelling capabilities and see if we can understand various factors affecting employee attrition on this IBM dataset using different analysis techniques.

Dataset

The Dataset consists of 35 columns. We Identified 8 Categorical, 8 Ordinals and rest all numeric. The dataset had an uneven distribution for attrition, so we had to balance the data by oversampling.



"age": Age	"HrRate": Hourly Rate	"PerfRate": Performance Rate
"Attr" : Attrition [Yes/No]	"JobInvl": Job Involvement	"RelnSat": Relationship Satisfaction
"BusTrvl" : Business Travel	"JobLvl": Job Level	"StkOpLvl": Stock Option Level
"DlyRate" : Daily Rate	"JobRole": Job Role	"TotWrkYr": Total Work Hours
"Dept": Department	"JobSat": Job Satisfaction	"TrnTimeLastYr": Training Time Last Year
"DistFrmHm" : Distance from Home	"MrgStatus": Marital Status	"WrkLifeBal": Work Life Balance
"Edu" : Education	"MonthRate": Monthly Rate	"YrsAtComp": Years at company
"EduField": Education Field	"NumCompWrk": No.of Company worked	"YrsInCurRole": Years in Current Role
"EnvSat": Environment Satisfaction	"OverTime": Over Time	"YrsLstProm": Years Last Promotion
"Gender": Gender [Male/Female]	"PercSalHike": Percentage Salary Hike	"YrsCurMngr": Years Current Manager
"EmployeeNumber": Employee Number	"Over18" : Age of Employee over 18	"MonthInc": Monthly Income
"EmployeeCount" : Total Employee	"StandardHours": Total Fixed Hours	

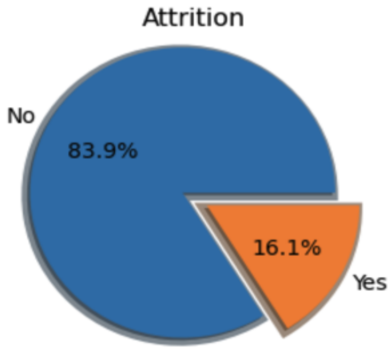


Figure 1: Dataset and Imbalance Graph

Data Exploration and Transformation

The data set was explored with some visualization techniques to analyze the behavior of attrition with other variables, and we see that attrition rate is high among people who are single also among people with a work life balance level (3: Better). We do also see that attrition rate is higher among people who does overtime. The data also reflected signs of attrition being higher among people who has low level of income.

In order to proceed with the analysis, we had to drop couple of variables which had no effect on attrition, and we applied log transformation to monthly income and preprocessing techniques to have even distribution among all variables.

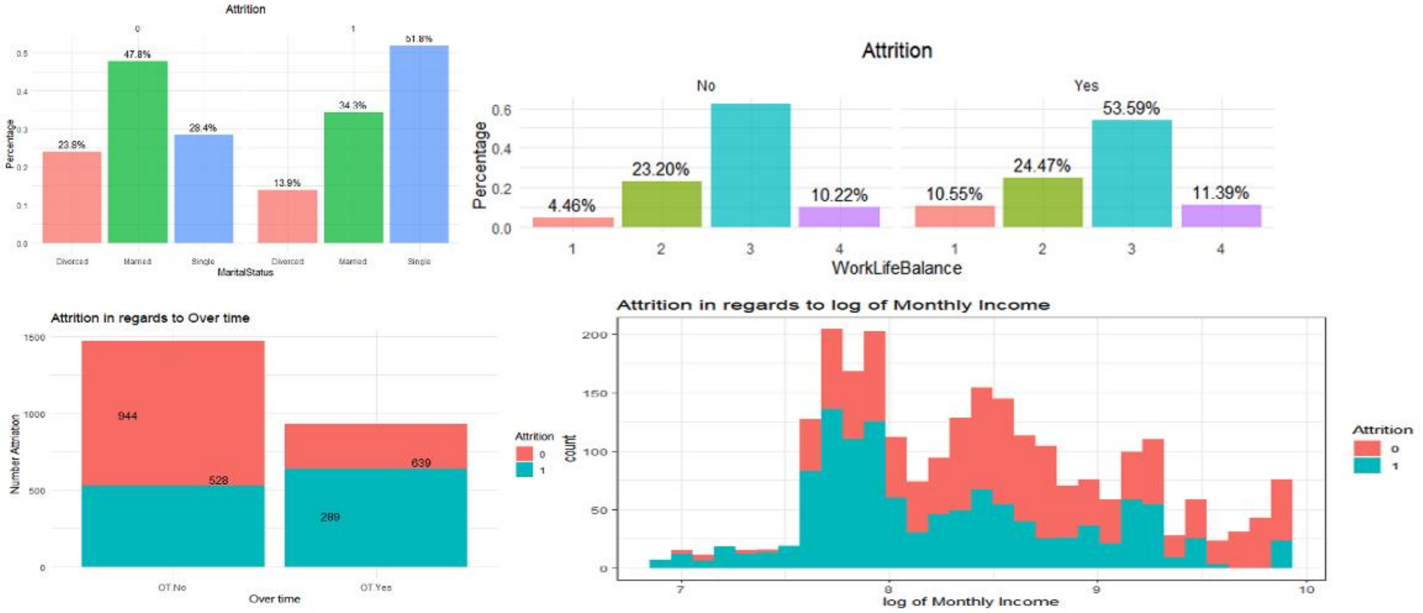


Figure 2: Exploratory Graphs

Analysis Techniques

In order to proceed with our analysis, we performed the below lines of analysis to infer the data and reach a consensus at the end. The different techniques that were used are as follows.

1. Principal Factor Analysis. (PFA)
2. Multiple Correspondence Analysis. (MCA)
3. Regression Technique. (RT)

We performed PFA to reduce the number of variables we would have to use in further lines of analysis by creating factors that would further explain how the variables were interacting. The results of the PFA, when using the dummy variables showed that we should have 13 factors which accounted for 58% of the variance in the data. These showed clear groupings that we could assign descriptive names and could use later in the regression analysis.

We then performed MCA to find relationships of different categorical variable to attrition and relationships among those categorical variables. The focus is on understanding how these variables are related to the likelihood of attrition occurring. By examining these relationships, we gained insights into the factors that might influence an individual's decision to leave or stay within the organization. According to MCA analysis we saw Travel frequently, overtime and people in sales department with lower education level and who are single are likely to have a higher frequency of Attrition.

We then took the 13 factors from our PFA analysis and performed RT to generate a predictive model with various performance measures. We performed several RT to find the optimal model and concluded that people who are doing Overtime, are in sales and marketing department also people with HR Profile and people who are single are likely to have a higher frequency of Attrition.

Technical Summary

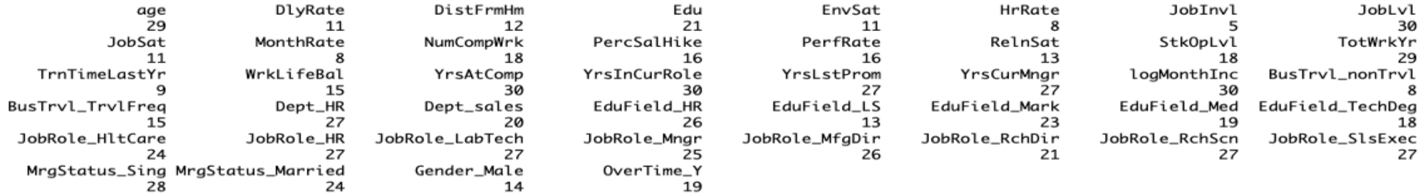
The different lines of techniques that were used for the analysis are explained below.

Principal Factor Analysis (PFA)

As the first line of analysis, we decided to use principal factor analysis to further analyze the interaction between our variables and identify any latent factors in our data. There are two main reasons for our use of principal factor analysis during our investigation. The first was to use factors as an exploratory technique to see how our variables were interacting with each other and how they could be grouped together. The second was to use the factors as a dimensionality reduction technique and reduce the 35 variables down to make them easier to explain and use in the regression analysis.

The first part of conducting the factor analysis was to clean the data and ensure that the variable types would be correct for factor analysis. First, the labels were all renamed to make them easier to understand (which was done earlier in the pre-processing of the data). Next was to change any categorical variables to numeric ordinal variables or binary variables where it made sense. The variable of interest, attrition, was also removed. This left us with 44 variables, including the dummies to perform the factor analysis with.

The next step was to perform a correlation test to see if there were any variables that were not correlated or were correlated with all the other variables. This was done using a P value of .05 and showed that we did not have any uncorrelated variables.



```
             age         DlyRate        DistFrmHm             Edu          EnvSat          HrRate         JobInvl           JobLvl
              29              11               12              21              11               8               5               30
          JobSat       MonthRate       NumCompWrk      PercSalHike        PerfRate        RelnSat        StkOpLvl         TotWrkYr
              11               8               18              16              16              13              18               29
  TrnTimeLastYr       WrkLifeBal        YrsAtComp     YrsInCurRole      YrsLstProm      YrsCurMngr    logMonthInc   BusTrvl_nonTrvl
               9              15               30              30              27              27              30                8
BusTrvl_TrvlFreq         Dept_HR       Dept_sales     EduField_HR     EduField_LS   EduField_Mark    EduField_Med EduField_TechDeg
              15              27               20              26              13              23              19               18
 JobRole_HltCare      JobRole_HR  JobRole_LabTech    JobRole_Mngr  JobRole_MfgDir  JobRole_RchDir  JobRole_RchScn  JobRole_SlsExec
              24              27               27              25              26              21              27               27
  MrgStatus_Sing MrgStatus_Married     Gender_Male      OverTime_Y
              28              24               14              19
```

Figure 3: Correlation test

We then created a correlation plot of our variables. This showed some likely factors such as all the variables relating to the years worked in a position, the performance and raises, and job level, age, and income.

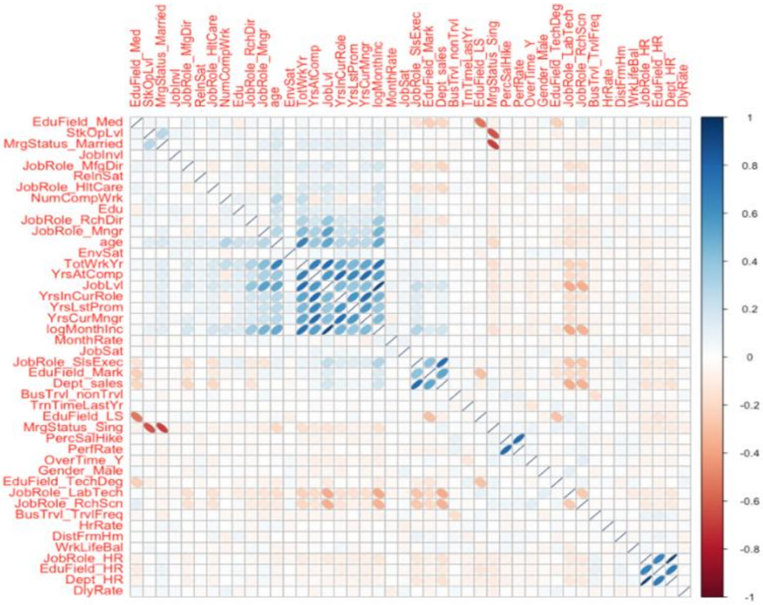


Figure 4: Correlation Plot

Next, we performed an initial PCA using the prcomp function to get an idea of the number of factors that we should expect. The results of the PCA were plotted on a scree graph to analyze. The n.obs were set to 22 variables to make it easier to read. The results showed that we should have around 18 factors, but the scree plot did not have a clear "knee". We also conducted parallel analysis was done to try and determine the correct number of factors to use. This suggested that we should use 18 factors.

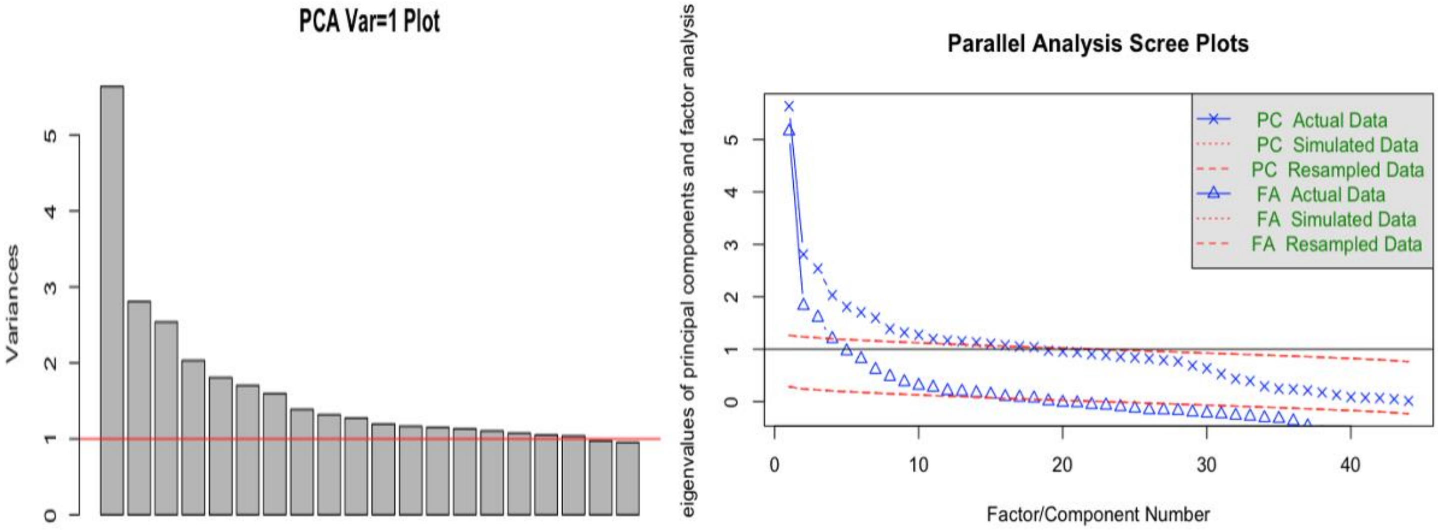
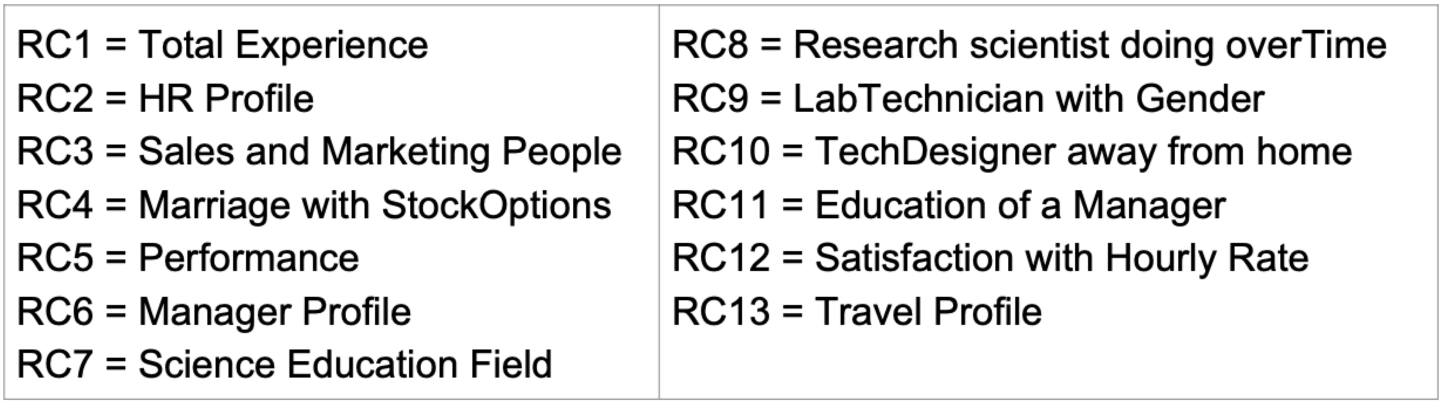


Figure 5: PCA & Parallel plots

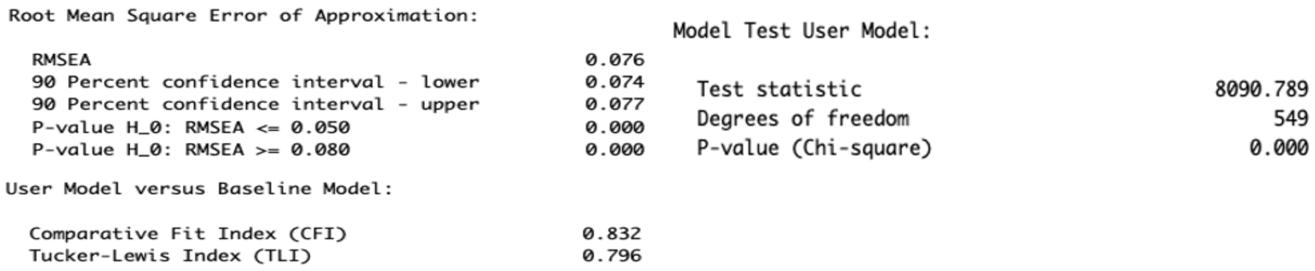
We next ran the PFA and realized that 18 factors would be too many as it was separating the variables into individual factors. We began reducing the number of factors and found 13 provided the best separation while still getting just under 60% of the variance. The factors can be broken down into the following categories:



RC1 = Total Experience	RC8 = Research scientist doing overTime
RC2 = HR Profile	RC9 = LabTechnician with Gender
RC3 = Sales and Marketing People	RC10 = TechDesigner away from home
RC4 = Marriage with StockOptions	RC11 = Education of a Manager
RC5 = Performance	RC12 = Satisfaction with Hourly Rate
RC6 = Manager Profile	RC13 = Travel Profile
RC7 = Science Education Field	

Confirmatory Factor Analysis (CFA)

In order to provide the diagnostics for the chosen set of factors, we performed a Confirmatory Factor analysis and from this we see that the RMSEA is 0.076 and the p value is 0 it provides strong evidence that the hypothesized model is a very good fit for the data and significantly better than a null model. This is also supported by the TLI and CFI score as both are close to 1.



```
Root Mean Square Error of Approximation:

  RMSEA                                          0.076
  90 Percent confidence interval - lower         0.074
  90 Percent confidence interval - upper         0.077
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.832
  Tucker-Lewis Index (TLI)                       0.796
```

```
Model Test User Model:

  Test statistic                              8090.789
  Degrees of freedom                               549
  P-value (Chi-square)                           0.000
```

Multiple Correspondence Analysis (MCA)

Multiple Correspondence Analysis (MCA) is a statistical technique used to analyze and visualize relationships between categorical variables in a data set. The MCA will transform the categorical variables into a lower-dimensional space, allowing for a visual representation of their similarities and associations.

In addition, this approach enables a comprehensive exploration of the underlying patterns and provides a valuable visual representation that will help identify variables that exhibit similar patterns and uncover important factors influencing employee attrition.

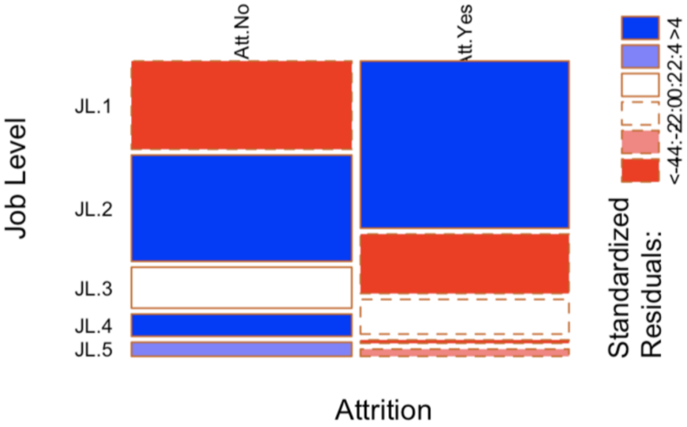


Figure 6: Mosaic plot(Attrition vs JobLevel)

Mosaic Plot

A mosaic plot on the left is a graphical representation that visualizes the relationship between two or more categorical variables.

The observed mosaic plot suggests a relationship between Job Level and the likelihood of attrition occurring. This finding indicates that employees with lower job levels may have a higher tendency to experience attrition than those with higher job levels.

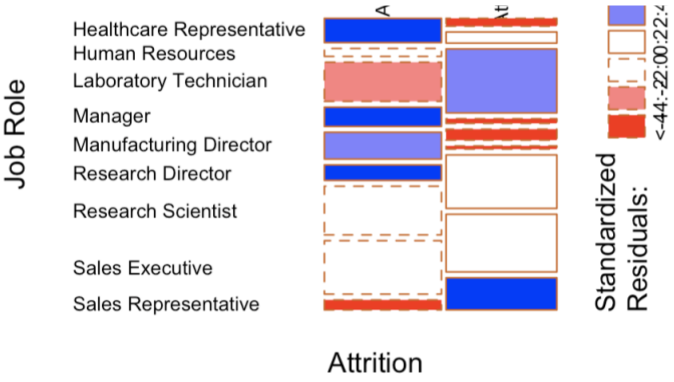


Figure 7: Mosaic plot(Attrition vs JobRole)

The figure on the left indicates that Sales Representatives and Laboratory Technicians have a higher tendency to experience attrition compared to other job roles. This suggests that these specific job roles may be associated with a higher attrition rate within the dataset.

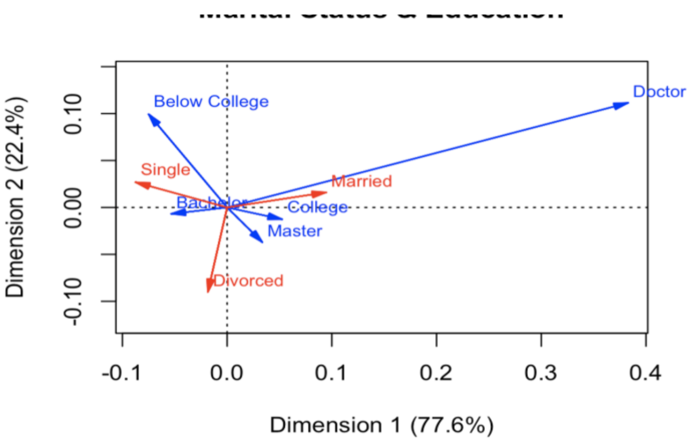


Figure 8: CA biPlot(MaritalStatus Vs Education)

Correspondence Analysis

Correspondence analysis (CA) is a multivariate statistical technique used to analyze categorical data and explore relationships between variables. It is often employed to identify patterns and associations in large contingency tables with multiple rows and columns.

In Figure 8. It suggests that in the dataset, the variable "Single" is closely associated with the education levels "Below College" and "bachelor’s degree."

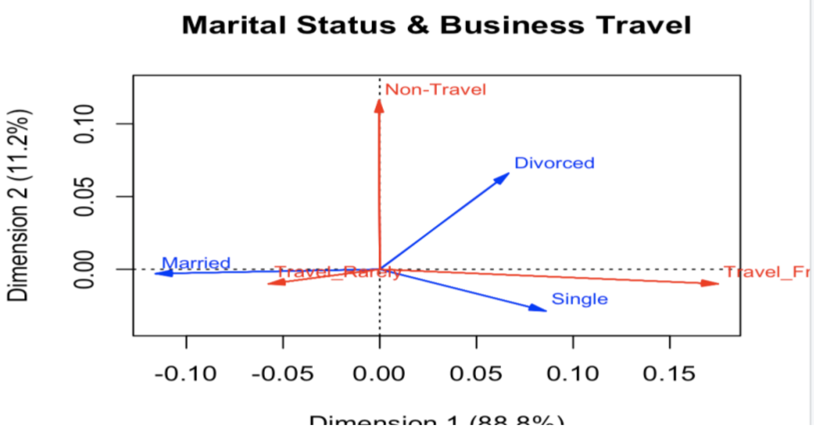


Figure 9: CA biplot (Marital Status Vs Buiss Trvl)

In Figure 9., it appears that there is a close relationship between the variables "Marital Status" and "Business Travel" in the analyzed dataset. Specifically, the analysis suggests that the category "Single" is closely associated with the category "Travel Frequently" in terms of business travel.

The proximity of these categories on the bi-plot indicates a strong relationship or association between them.

Comparing Attrition to six categorical variables in MCA Plot A, it appears that top three relating to Attrition is being Single, working overtime, and travel frequently.

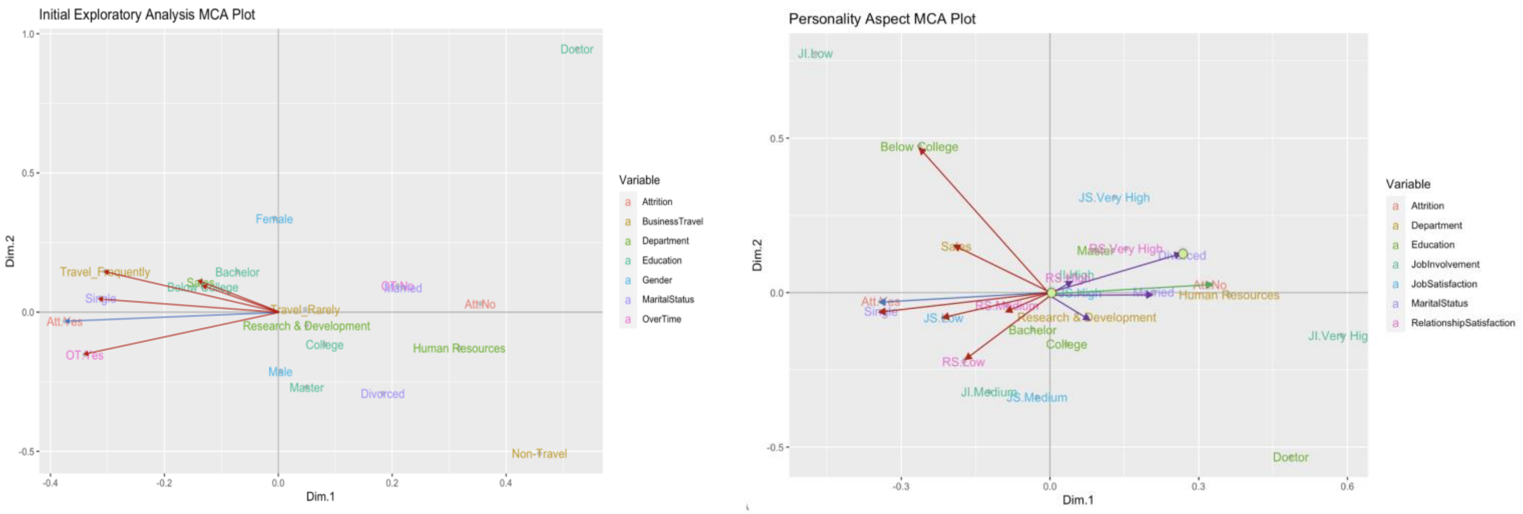


Figure 10: MCA Plot A & B

The MCA plot B suggests an association between individual personality aspects and attrition in the analyzed dataset. Specifically, the bi-plot indicates that individuals at the early stages of their careers are more likely to experience attrition. Furthermore, the bi-plot suggests that mature individuals with a clearer understanding of their career path are less likely to experience attrition. This may be attributed to their knowledge and experience, allowing them to make more informed career decisions. Additionally, individuals with family responsibilities, such as children and a spouse, may be less willing to travel for work, potentially reducing their likelihood of attrition.

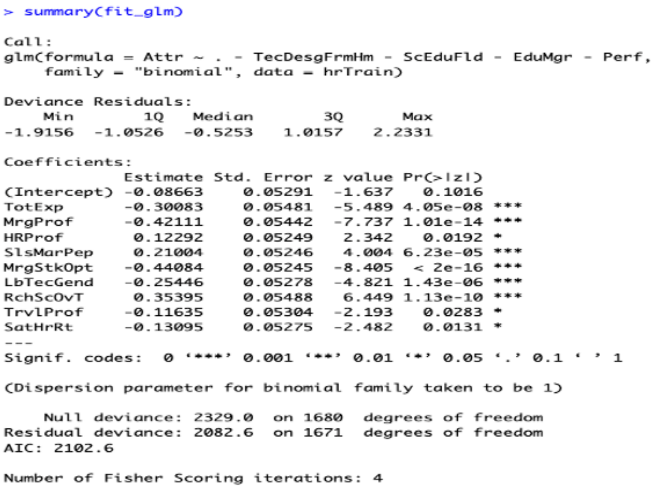
Regression Techniques (RT)

The different regression techniques that were used for the analysis are:

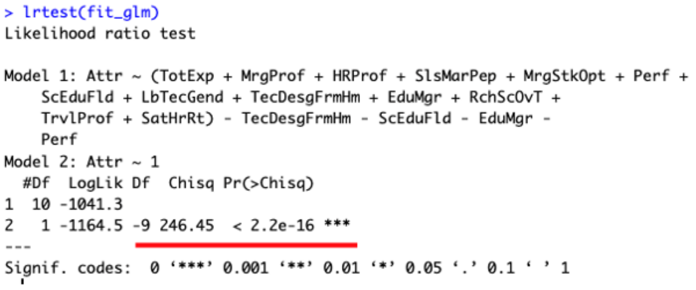
- Logistic Regression
- Regsubset Regression
- Ridge Regression
- Lasso Regression
- Eslasticnet Regression
- Relaxed Lasso Regression

Logistic Regression

- We performed a logistic regression dividing data into 70% training and 30% testing with manual variable selection and measured the performance.
- The model turned out decent with selected variable being significant at $p < 0.05$ having a goodness of fit of chi square = 16.73, $p < 2.2e-16$ ***.



```
> summary(fit_glm)

Call:
glm(formula = Attr ~ . - TecDesgFrmHm - ScEduFld - EduMgr - Perf,
    family = "binomial", data = hrTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9156  -1.0526  -0.5253   1.0157   2.2331

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08663    0.05291  -1.637   0.1016
TotExp      -0.30083    0.05481  -5.489 4.05e-08 ***
MrgProf     -0.42111    0.05442  -7.737 1.01e-14 ***
HRProf       0.12292    0.05249   2.342   0.0192 *
SlsMarPep    0.21004    0.05246   4.004 6.23e-05 ***
MrgStkOpt   -0.44084    0.05245  -8.405  < 2e-16 ***
LbTecGend   -0.25446    0.05278  -4.821 1.43e-06 ***
RchScOvT     0.35395    0.05488   6.449 1.13e-10 ***
TrvlProf    -0.11635    0.05304  -2.193   0.0283 *
SatHrRt     -0.13095    0.05275  -2.482   0.0131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2329.0  on 1680  degrees of freedom
Residual deviance: 2082.6  on 1671  degrees of freedom
AIC: 2102.6

Number of Fisher Scoring iterations: 4
```



```
> lrtest(fit_glm)
Likelihood ratio test

Model 1: Attr ~ (TotExp + MrgProf + HRProf + SlsMarPep + MrgStkOpt + Perf +
    ScEduFld + LbTecGend + TecDesgFrmHm + EduMgr + RchScOvT +
    TrvlProf + SatHrRt) - TecDesgFrmHm - ScEduFld - EduMgr -
    Perf
Model 2: Attr ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -1041.3
2   1 -1164.5 -9 246.45  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

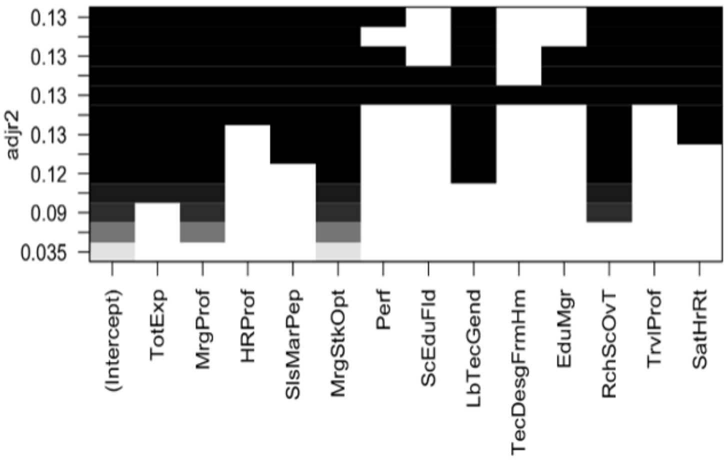
Model Equation : log(p/(1-p)) = -0.08 - 0.42(MrgProf) -0.30(TotExp) + 0.21(SlsMarPep) + 0.12(HRProf) - 0.44(MrgStkOpt) - 0.25(LbTecGend) + 0.35(RchScOvT) - 0.11(TrvlProf) - 0.13(SatHrRt)

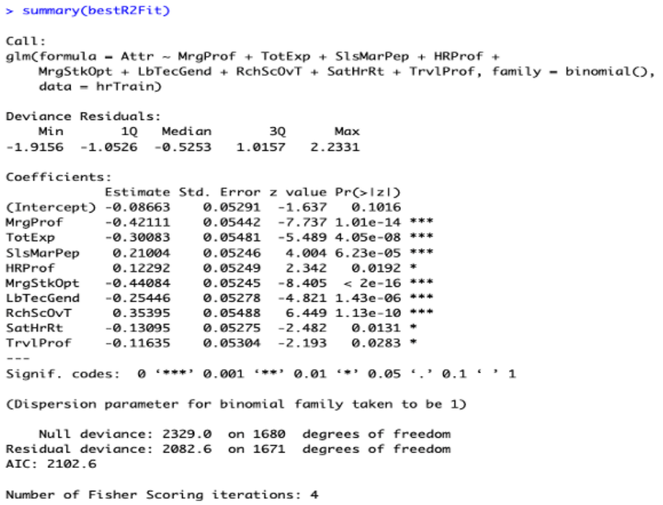
Figure 11: Logistic Regression summary

- Since the data was balanced with oversampling the threshold was set at 0.5 to measure performance (fig. 17).
- And we found no significant difference to indicate overfitting.

Regsubset Regression

- We performed a Regsubset regression using “adjR2” and “bic” scale to measure the performance and found out the model with adjR2 did better after removing the non-significant variables and the performance (fig. 17) was same as logistic regression and there were no signs of overfitting.





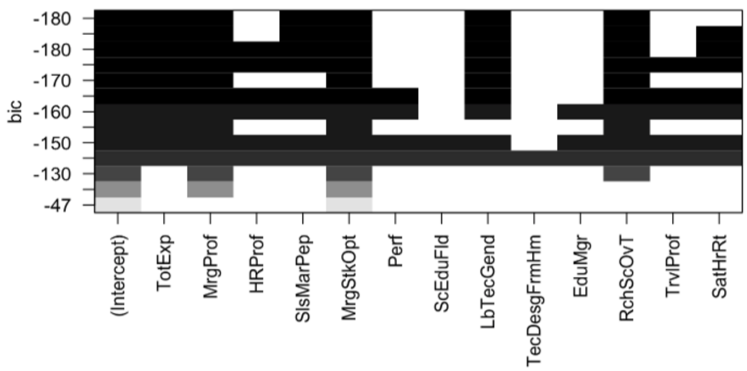
```
> summary(bestR2Fit)

Call:
glm(formula = Attr ~ MrgProf + TotExp + SlsMarPep + HRProf +
    MrgStkOpt + LbTecGend + RchScOvT + SatHrRt + TrvlProf, family = binomial(),
    data = hrTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9156  -1.0526  -0.5253   1.0157   2.2331

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08663    0.05291  -1.637   0.1016
MrgProf     -0.42111    0.05442  -7.737 1.01e-14 ***
TotExp      -0.30083    0.05481  -5.489 4.05e-08 ***
SlsMarPep    0.21004    0.05246   4.004 6.23e-05 ***
HRProf       0.12292    0.05249   2.342   0.0192 *
MrgStkOpt   -0.44084    0.05245  -8.405  < 2e-16 ***
LbTecGend   -0.25446    0.05278  -4.821 1.43e-06 ***
RchScOvT     0.35395    0.05488   6.449 1.13e-10 ***
SatHrRt     -0.13095    0.05275  -2.482   0.0131 *
TrvlProf    -0.11635    0.05304  -2.193   0.0283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2329.0  on 1680  degrees of freedom
Residual deviance: 2082.6  on 1671  degrees of freedom
AIC: 2102.6

Number of Fisher Scoring iterations: 4
```

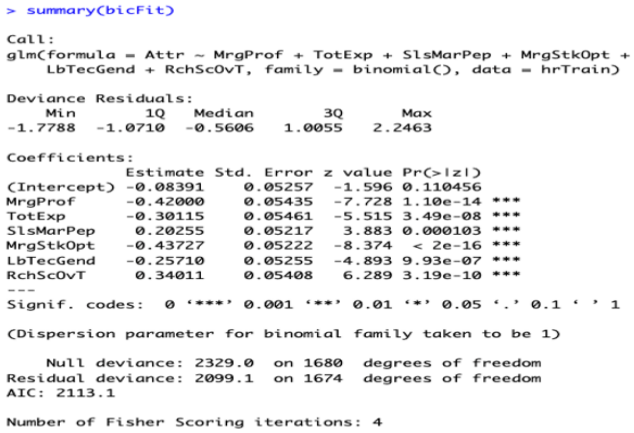
Figure 12: Regsubset Regression summary

Regsubset(adjR2) regression model equation:

log(p/(1-p)) = - 0.08 - 0.42(MrgProf) - 0.30(TotExp) + 0.21(SlsMarPep) + 0.12(HRProf) - 0.44(MrgStkOpt) - 0.25(LbTecGend) + 0.35(RchScOvT) - 0.11(TrvlProf) - 0.13(SatHrRt)

bic criterion technique





```
> summary(bicFit)

Call:
glm(formula = Attr ~ MrgProf + TotExp + SlsMarPep + MrgStkOpt +
    LbTecGend + RchScOvT, family = binomial(), data = hrTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7788  -1.0710  -0.5606   1.0055   2.2463

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08391    0.05257  -1.596 0.110456
MrgProf     -0.42000    0.05435  -7.728 1.10e-14 ***
TotExp      -0.30115    0.05461  -5.515 3.49e-08 ***
SlsMarPep    0.20255    0.05217   3.883 0.000103 ***
MrgStkOpt   -0.43727    0.05222  -8.374  < 2e-16 ***
LbTecGend   -0.25710    0.05255  -4.893 9.93e-07 ***
RchScOvT     0.34011    0.05408   6.289 3.19e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2329.0  on 1680  degrees of freedom
Residual deviance: 2099.1  on 1674  degrees of freedom
AIC: 2113.1

Number of Fisher Scoring iterations: 4
```

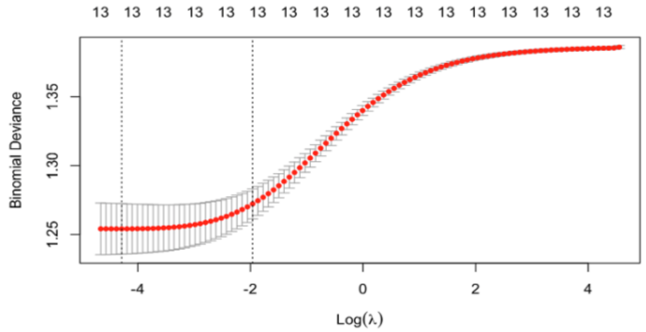
Figure 13: bic summary

Regsubset(bic) regression model equation:

log(p/(1-p)) = -0.08 - 0.42(MrgProf) -0.30(TotExp) + 0.20(SlsMarPep) - 0.44(MrgStkOpt) - 0.25(LbTecGend) + 0.34(RchScOvT)

Ridge Regression

We performed ridge regression as another technique to verify the performance, we select lambda.min for our lambda parameter as we want to have maximum confidence interval to capture most of the true positives. The threshold was again set at 0.5 to measure performance (fig. 17). We did not find any significant difference to indicate overfitting.



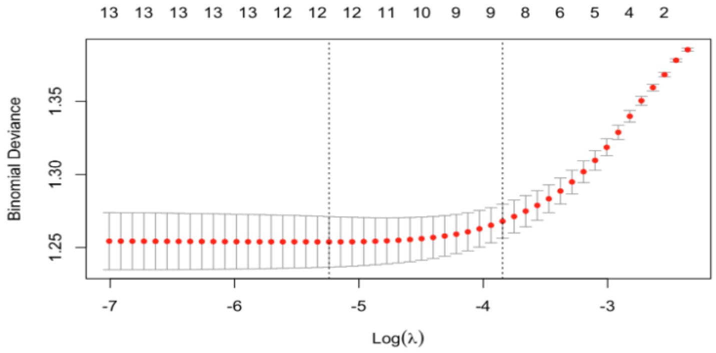
Ridge Equation:

log(p/(1-p)) = - 0.082 - 0.39(MrgProf) - 0.27(TotExp) + 0.20(SlsMarPep) + 0.12(HRProf) - 0.41(MrgStkOpt) - 0.05(Perf) + 0.03(ScEduFld) - 0.23(LbTecGend) + 0.32(RchScOvT) - 0.04(EduMgr) - 0.01(TecDesgFrmHm) - 0.11(SatHrRt) - 0.10(TrvlProf)

Figure 14: Ridge Plot

Lasso Regression

We performed a lasso regression to verify the performance again and it performed almost similar. We select lambda.min for our lambda parameter as we want to have maximum confidence interval to capture most of the true positives. The threshold was set at 0.5 to measure the performance (fig. 17).



Lasso Model Equation:

log(p/(1-p)) = - 0.08 - 0.39(MrgProf) - 0.27(TotExp) + 0.18(SlsMarPep) + 0.10(HRProf) - 0.41(MrgStkOpt) - 0.03(Perf) - 0.008(ScEduFld) - 0.23(LbTecGend) + 0.32(RchScOvT) - 0.02(EduMgr) - 0.10(SatHrRt) + 0.09(TrvlProf)

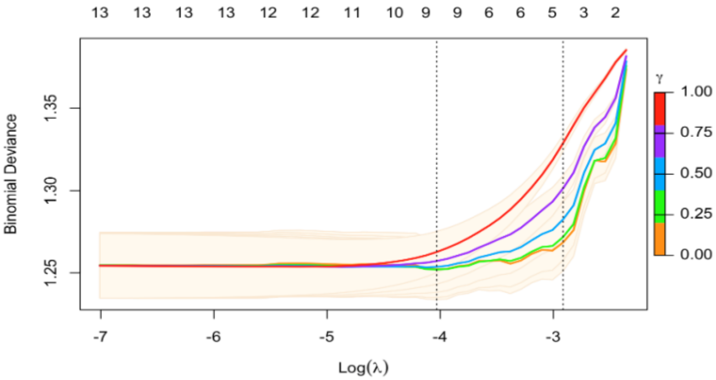
Figure 15: Lasso Plot

Elasticnet Regression

For elasticnet we ran a for loop with a range of alpha 0 to 1 with an increment of 0.01 and captured the performance. And it reported that model at alpha = 0.55 did better with performance which we see is same as the performance of lasso. The threshold was set at 0.5 to measure the performance (fig. 17).

Relaxed Lasso

In order to find a parsimonious model without sacrificing much of performance (fig. 17) we performed the relaxed lasso.



Relaxed Lasso Equation:

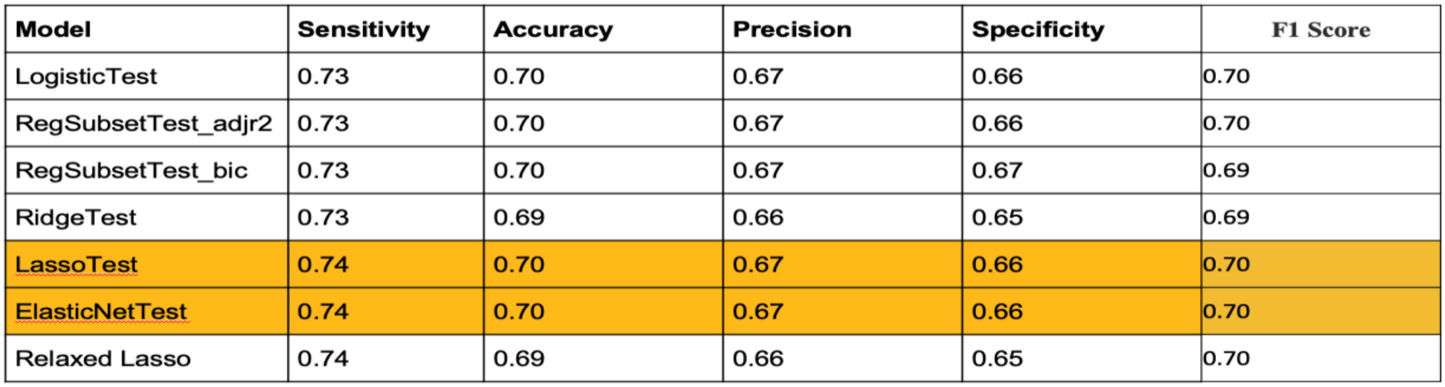
log(p/(1-p)) = - 0.08 - 0.42(MrgProf) - 0.30(TotExp) + 0.21(SlsMarPep) + 0.12(HRProf) - 0.44(MrgStkOpt) - 0.25(LbTecGend) + 0.35(RchScOvT) - 0.13(SatHrRt) - 0.11(TrvlProf)

Figure 16: Relaxed Lasso Plot

From the plot we can see that gamma at 0 gives us the best fit which has the least penalization and created a more parsimonious model removing 3 variables and sacrificing 1% in model accuracy and specificity.

Model Comparison:

From the Performance matrix we see Lasso is doing slightly better than all other model which is also reported in elastic net with similar performance. The Relaxed lasso however gave us a parsimonious model, but we wanted to capture most of the sensitivity and accuracy of all the model.



Model	Sensitivity	Accuracy	Precision	Specificity	F1 Score
LogisticTest	0.73	0.70	0.67	0.66	0.70
RegSubsetTest_adjr2	0.73	0.70	0.67	0.66	0.70
RegSubsetTest_bic	0.73	0.70	0.67	0.67	0.69
RidgeTest	0.73	0.69	0.66	0.65	0.69
LassoTest	0.74	0.70	0.67	0.66	0.70
ElasticNetTest	0.74	0.70	0.67	0.66	0.70
Relaxed Lasso	0.74	0.69	0.66	0.65	0.70

Figure 17: Model Comparison table

So, if we look at the model equation for Lasso, we have some of the highest coefficients for **RchScOvT** which is a factor of Overtime With research scientist, **SlsMarPep** which is factor of Sales and marketing people, **HRProf** which is a factor of HR Profile, **MrgProf** affecting negatively which is a factor of Manager Profile, **MrgStkOpt** affecting negatively which is a factor of Marriage and stock option. These all factor has a significant influence on the attrition which is kind of supporting our initial hypothesis.

Conclusion

- The PFA was able to effectively reduce the number of variables and create logical groupings form them. This allowed us to create a more parsimonious model to use in the later regression.

- According to MCA analysis we saw that individuals at the early stages of their careers are more likely to experience attrition. Furthermore, it suggested that mature individuals with a clearer understanding of their career path are less likely to experience attrition. Additionally, people who travel frequently, especially the sales department personnel, who are working overtime and are single has a higher frequency of attrition.

- According to Regression people who are doing Overtime, are in sales and marketing department also people with HR Profile and people who are single are likely to have a higher frequency of Attrition.

Summarizing the above observation, we reach to a consensus that people who travel frequently, do overtime and are in lower job band and are single are prone to attrition. And probably Companies need to take necessary steps to address/monitor these factors stringently to control attrition.

Punchlist Items Addressed

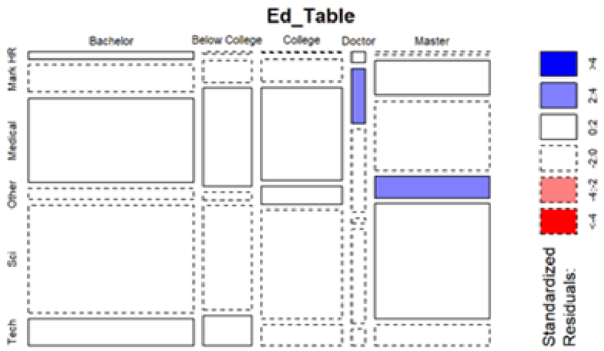
1. We revisited the PFA according to the feedback and included all variables having dummies and did the Scree plot and parallel analysis. Additionally, we also performed the CFA and reported the diagnostics. These factors were used in the regression analysis.
2. For the feedback w.r.t MCA we performed pairwise CA for multiple variables and is reported in the document. The Result of MCA corresponded to the rest of the analysis and has been highlighted in the conclusion. The MCA already has the dummy variables included and these are in the form of class levels in the category and the result coincides to the factor analysis and regression analysis.
3. As per the feedback for Regression we renamed the factors to meaningful names making model equations easier to interpret. We also performed relaxed lasso technique reported in the technical summary.

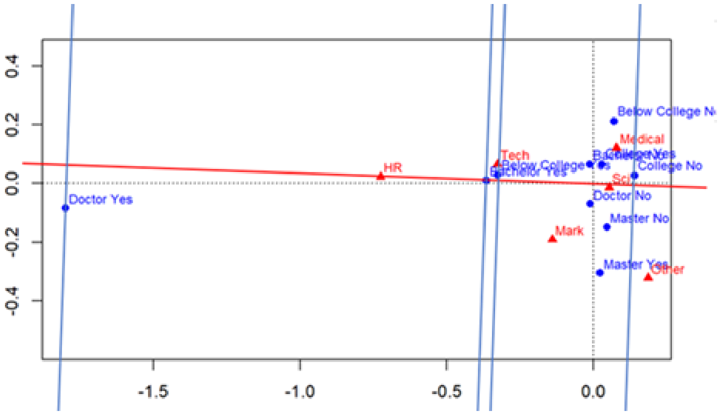
Appendix

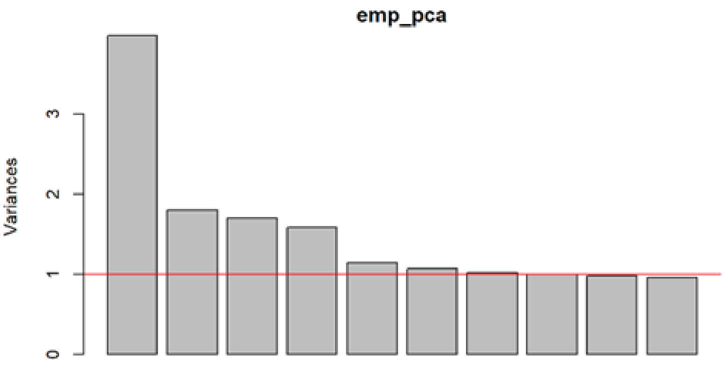
Individual Summary

– Zach Hollis

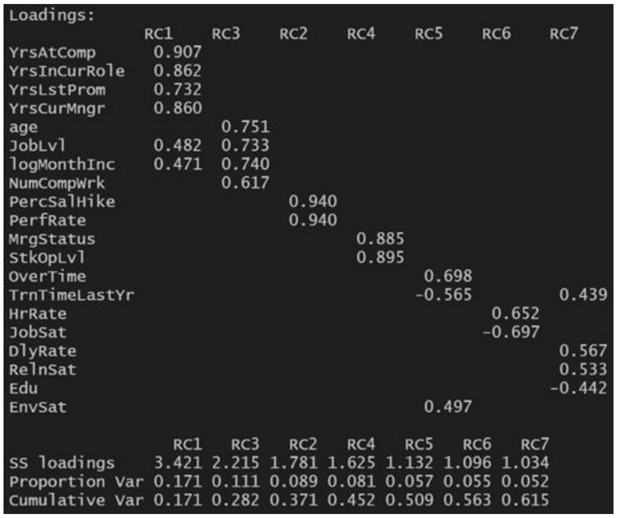
To start the analysis for our dataset, the group decided to all look at different variables and create some visualizations and conduct some exploratory analysis. I decided to look at the education variables for our dataset and see how they interacted with the attrition variable which was our variable of interest. Since these were categorical, I decided to try correspondence analysis and created mosaic plots for the education type and education level. The results did show some relationships between the level of education and education type as well as which education levels corresponded with attrition. After conducting this correspondence analysis, I decided to try and create a regression model using dummy variables for the categorical variables. However, these were not significant when predicting attrition but were significant when predicting monthly income. These models were useful for exploring the data, but we would create better models for the final report that better explained the data.







The next part of the project that I worked on was conducting principal factor analysis to try and uncover the relationships between our variables. I wanted to do this for two reasons. The first was to explore the relationships between the variables and see how they were interacting with each other. The second was for dimension reduction that could be used in the regression analysis that would Jiten perform for his part of the project. After preprocessing the data and removing unnecessary variables I started by running a correlation test to see if any variables were correlated with too many other variables or had no correlation at all. This showed that we had 7 variables that were not correlated with any other variables and could be removed. There were no variables that were correlated with all or close to all other variables. This left us with the 20 shown on the correlation plot here. We can see some clear grouping around the years worked variables, the performance and raise variables, and age, job level, and monthly income. The next step I took was to run prcomp to try and see how many factors we would need to have to capture enough variance. Using the V=1 method we can see that we could pick 7 PCs to use in the principal factor analysis. Unfortunately, the scree plot was less clear, but there could be a small knee after the 4th PC. To try and get a better idea of how many factors to use I next performed a parallel analysis. This suggested that the number of factors should be 6 and the number of components should be 5. Therefore, I decided that the number of factors would likely be 6 or 7 and would test each to see which returned the better result.



```
Loadings:
              RC1    RC3    RC2    RC4    RC5    RC6    RC7
YrsAtComp      0.907
YrsInCurRole   0.862
YrsLstProm     0.732
YrsCurMngr     0.860
age                   0.751
JobLvl         0.482  0.733
logMonthInc    0.471  0.740
NumCompWrk            0.617
PercSalHike                  0.940
PerfRate                     0.940
MrgStatus                           0.885
StkOpLvl                            0.895
OverTime                                   0.698
TrnTimeLastYr                             -0.565         0.439
HrRate                                            0.652
JobSat                                           -0.697
DlyRate                                                  0.567
RelnSat                                                  0.533
Edu                                                     -0.442
EnvSat                                     0.497

                 RC1   RC3   RC2   RC4   RC5   RC6   RC7
SS loadings    3.421 2.215 1.781 1.625 1.132 1.096 1.034
Proportion Var 0.171 0.111 0.089 0.081 0.057 0.055 0.052
Cumulative Var 0.171 0.282 0.371 0.452 0.509 0.563 0.615
```

I started by using 6 factors and ran a principal factor analysis using the principal function from the psych package. This did not return the best result as it left out some variables and only accounted for 56% of the cumulative variance in the data. However, it did start to show a clear grouping of correlated variables which made logical sense. The second principal factor analysis using 7 factors did a better job of separating the variables into clear groupings. It also accounted for 61% of the cumulative variance which cleared the 60% threshold for principal analysis. The first RC could be called "years of service" since it is predominantly made up of the variables relating to the amount of time spent in the company and in a certain position. The second factor (RC3) could be called "seniority" since it is the age and job level. The number of other companies worked at and higher income also makes sense since this would likely mean the employee has more experience. The next factor is clearly performance since it is the performance rate of the employee and the raises, they receive. Comparing these to the corrplot above also shows that they make sense as they are the same groupings. Marriage status and stock options also are strongly correlated in the plot and are grouped together here (perhaps implying that older employees with families are more likely to invest). The last three RCs are less correlated but still make sense when compared to the corrplot. These 7 factors would be used in the regression analysis to see how they would perform compared to other methods, such as using dummy variables. In addition to this other team members would use methods such as correspondence analysis and MCA to try and see how the categorical variables that I could not include in this factor analysis interacted.

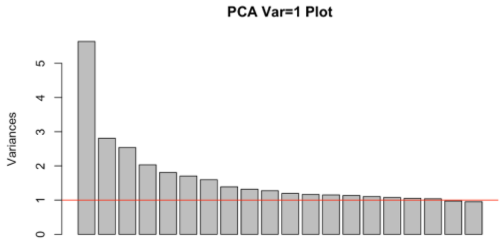
Overall, I think that this project was a great way of trying to combine all the different techniques for analysis that we have covered over the quarter. We started by making exploratory visualizations and transforming variables when needed. Then attempted to build a regression model to predict attrition and used elastic net, lasso, and ridge to compare different models. Over the course of the project our idea of what we should be doing changed as we continued to investigate the data. At first, we had thought that we would try and just build the best model for predicting attrition or another variable like income but as we progressed, we changed from that to trying to tell a story about what the data really showed and how the different variables were interacting. I think this showed what we had learned in the class well since it was not just about getting the best prediction rate. We could show how the factors are grouping together logical variables and get a better explanation of the data that makes sense in the real world. The data set we chose also helped since it contained numeric, ordinal, and categorical variables. This allowed us to use all the different techniques that we had learned during the course. I think this was a good way to show how all of these different types of analysis should be used together to get a better sense of what is actually happening in the data. In past courses, we had focused on simply getting the best prediction or r^2 value so this was a way to learn how to better analyze the underlying patterns in the data and be better at explaining what was going on beneath the surface.

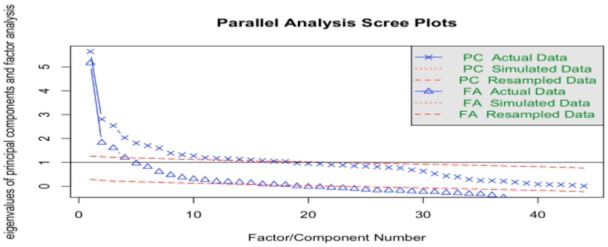
– Jiten Mishra

To start working on the selected dataset “IBM HR Attrition Dataset” we performed some explanatory analysis to identify the lines of analysis that we need to work on this dataset. Since the dataset is related to attrition of an employee our obvious target variable was attrition as a categorical variable. The Team did pick other variables as part of the first step of analyzing the data and I went ahead picking marital status with gender and overtime and analyzed the data how they affected attrition and reached to the initial hypotheses based on different visual plots that male employees who work overtime, are single have a higher attrition rate. While performing the initial analysis I also found out that people who potentially have a low income, have a higher attrition rate. Additionally, I also observed a significant attrition rate among young employees.

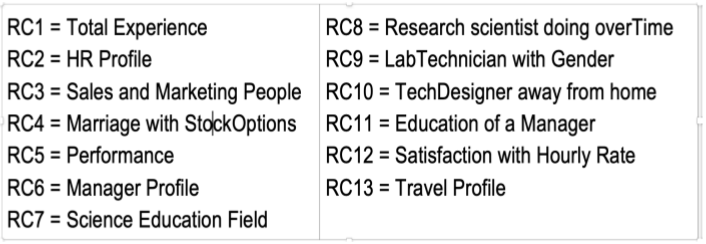
On collaborating with our initial analysis, we then went ahead on deciding our various lines of analysis that we need to perform on the data set and my part of analysis was Regression Analysis.

To proceed with the analysis, I initially did some preprocessing like Converting categorical data as factors and created dummies. Since we had Monthly Income as a variable, I log transformed it. Because of class imbalance in the data, I oversampled the data for our analysis. Since I intended on using full set of data for regression, I performed a initial correlation test and found no variable being uncorrelated to any other variable, I then performed PCA and PFA on this set of data to get necessary factors and identified 13 factors covering 58% of the variance that was needed to proceed with the regression techniques. To confirm I also performed the CFA as part of feedback with the help of my team together and measured the diagnostics.





As part of the feedback, before proceeding with the regression analysis I renamed the factors to meaningful names for better interpretability.


RC1 = Total Experience
RC2 = HR Profile
RC3 = Sales and Marketing People
RC4 = Marriage with StockOptions
RC5 = Performance
RC6 = Manager Profile
RC7 = Science Education Field
RC8 = Research scientist doing overTime
RC9 = LabTechnician with Gender
RC10 = TechDesigner away from home
RC11 = Education of a Manager
RC12 = Satisfaction with Hourly Rate
RC13 = Travel Profile

TotExp, HRProf, SlsMarPep, MrgStkOpt, Perf, MrgProf, ScEduFld, RchScOvT, LbTecGend, ,TecDesgFrmHm, EduMgr, SatHrRt , TrvlProf

Renamed the factors for interpretability.

And the following techniques was performed by me to predict the attrition class:

Logistic Regression: I performed logistic regression on the data, dividing it into training and testing sets. With manual variables selection and achieved a good model fit, with the selected variables being statistically significant and measured the performance of the model.

RegSubset Regression: I used regsubset regression with adjusted R-squared and BIC (Bayesian Information Criterion) to select the best model and found that the model with adjusted R-squared performed better after removing non-significant variables and the performance of the model was measured.

Ridge Regression: I performed Ridge regression as another technique to verify the performance of the model. I selected lambda.min as the lambda parameter to have a maximum confidence interval. I generated the model at lambda min and measured the model performance.

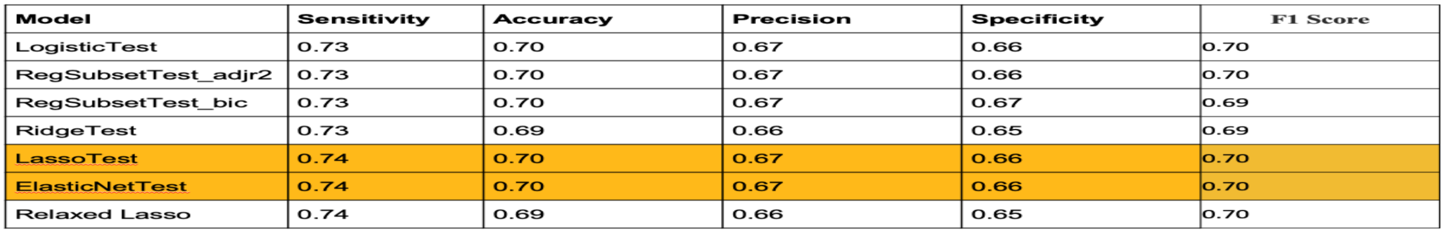
Lasso Regression: Lasso regression was performed to further validate the performance of the model, and it showed similar results to the ridge regression as per the model performance.

Elastic Net Regression: I ran a for loop with a range of alpha values (0 to 1 with an increment of 0.01) for elastic net regression. And found that the model at alpha = 0.55 performed the best, with results similar to lasso regression.

Relaxed Lasso: As part of the feedback, we also performed relaxed lasso with an intention to get a parsimonious model having a good performance w.r.t predicting attrition. This part was done by me and Denvir together and me measured the performance of model.

While performing the above techniques I made sure and kept note that I don't see any multicollinearity issues and kept measuring the goodness of fit at each step of the technique.

The performance of the models was compared using sensitivity, accuracy, precision, specificity and F1 score metrics. From the analysis I concluded that the lasso regression model performed slightly better than the others, including elastic net regression reporting the same with a sensitivity of 74%, accuracy of 70%, precision 67%, specificity 66% and F1 score of 70%.



Model	Sensitivity	Accuracy	Precision	Specificity	F1 Score
LogisticTest	0.73	0.70	0.67	0.66	0.70
RegSubsetTest_adjr2	0.73	0.70	0.67	0.66	0.70
RegSubsetTest_bic	0.73	0.70	0.67	0.67	0.69
RidgeTest	0.73	0.69	0.66	0.65	0.69
LassoTest	0.74	0.70	0.67	0.66	0.70
ElasticNetTest	0.74	0.70	0.67	0.66	0.70
Relaxed Lasso	0.74	0.69	0.66	0.65	0.70

Based on the Lasso model equations, I concluded that we have some of the highest coefficients for RchScOvT which is a factor of Overtime With research scientist, SlsMarPep which is factor of Sales and marketing people, HRProf which is a factor of HR Profile, MrgProf affecting negatively which is a factor of Manager Profile, MrgStkOpt affecting negatively which is a factor of Marriage and stock option. These all factors have a significant influence on the attrition which is kind of supporting our initial hypothesis.

Other techniques that were performed by the team was MCA which was a new technique to learn, the team also performed LDA which was interesting to analyze but it was later dropped from the final report as it did not show strong results w.r.t analyzing the data. A different approach of PFA was also performed by the team which was interesting but was modified based on feedback to stay in line with the analysis.

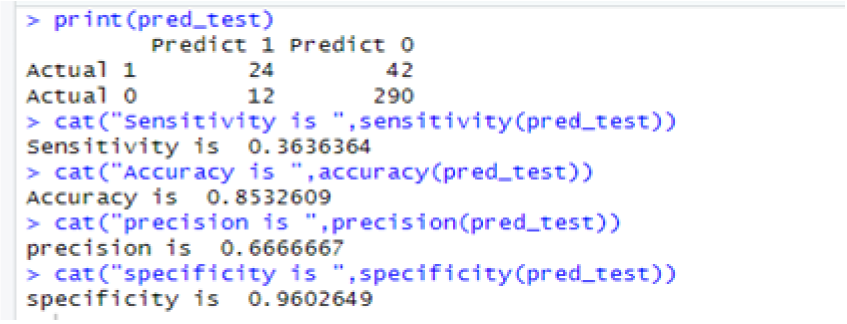
The team did group together to complete the final report and everyone was cooperative and helpful performing their part of the job and finally the teamwork was fruitful.

Overall, this project was a great learning process in implementing almost all the techniques that we learned in the class. As the team also implemented a new line of analysis in this project, I did get to learn new stuff that was not part of the course. With all the quizzes and assignments that were assigned to me I had a great learning curve and it helped me gain more in-depth knowledge of the course. From the prior course of DSC 423 I did learn the basics of getting a working model for any data set and this advanced course helped me understand detailed techniques and ways of getting a better and parsimonious model for the same dataset and analyzing the data in various techniques.

– Denvir Gama

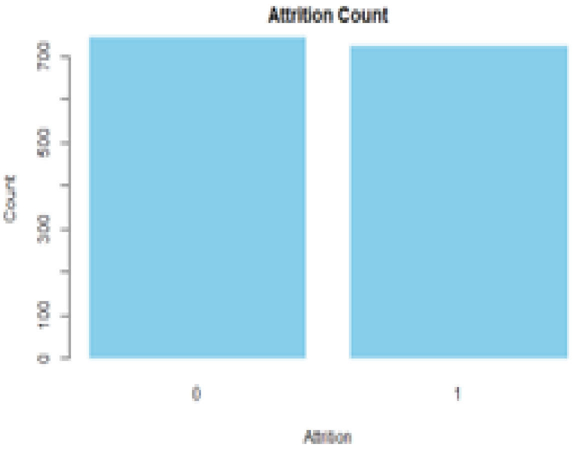
LDA:

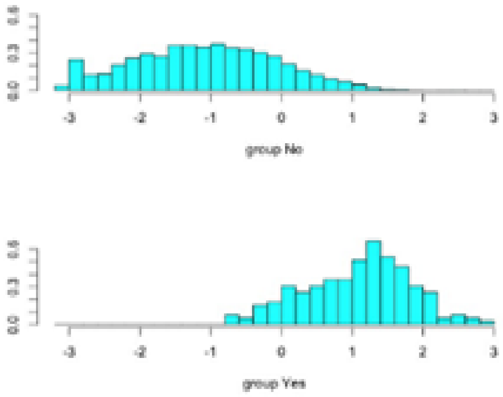
In my application of Linear Discriminant Analysis (LDA), I employed a comprehensive set of categorical predictors to predict the variable of attrition. By including all relevant categorical predictors in my analysis, I aimed to capture the potential influence of various factors on the attrition outcome. LDA, a widely used statistical technique, allowed me to model and analyze the relationship between these categorical predictors and the attrition variable, ultimately providing valuable insights into the factors associated with attrition. In the initial phase of my analysis, I encountered an unbalanced dataset, which resulted in certain challenges when examining the model's plot. The plot revealed instances where data points appeared to overlap, indicating potential ambiguity in distinguishing between different classes. This overlapping phenomenon can be attributed to the inherent nature of the unbalanced dataset, where one class (e.g., attrition level "yes") is significantly underrepresented compared to the other class. This class imbalance posed a particular concern when assessing the model's performance metrics. Notably, the sensitivity metric, which measures the ability of the model to correctly identify instances of attrition, yielded a relatively low value of approximately 0.36. This indicates that the model exhibited limited effectiveness in capturing and correctly predicting instances of attrition with a "yes" label. Consequently, the model's performance suffered from a reduced ability to identify and classify attrition cases accurately due to the scarce representation of the attrition "yes" class in the unbalanced dataset.

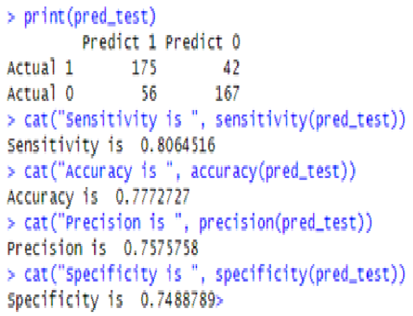


```
> print(pred_test)
         Predict 1 Predict 0
Actual 1        24        42
Actual 0        12       290
> cat("Sensitivity is ",sensitivity(pred_test))
Sensitivity is  0.3636364
> cat("Accuracy is ",accuracy(pred_test))
Accuracy is  0.8532609
> cat("precision is ",precision(pred_test))
precision is  0.6666667
> cat("specificity is ",specificity(pred_test))
specificity is  0.9602649
```

In order to address the initial issue of class imbalance in the dataset, I employed the Rose library to perform oversampling. This technique helped to rectify the unequal representation of "Yes" and "No" attrition instances, leading to a more balanced distribution as depicted in the bar graph. The subsequent implementation of LDA on the oversampled data showcased significant improvements in model performance compared to the previous analysis. The LDA plot illustrated enhanced separability between the attrition classes, indicating better discrimination after the dataset was balanced.







```
> print(pred_test)
         Predict 1 Predict 0
Actual 1       175        42
Actual 0        56       167
> cat("Sensitivity is ", sensitivity(pred_test))
Sensitivity is  0.8064516
> cat("Accuracy is ", accuracy(pred_test))
Accuracy is  0.7772727
> cat("Precision is ", precision(pred_test))
Precision is  0.7575758
> cat("Specificity is ", specificity(pred_test))
Specificity is  0.7488789>
```

Ordinal Logistic Regression:

Utilizing monthly income as the target variable in an ordinal logistic regression enables the prediction of income levels within the HR dataset. This modeling approach offers valuable insights into the factors that exert influence on income and facilitates the identification of patterns and variables that contribute to higher or lower income brackets. Additionally, by running an ordinal logistic regression, it becomes possible to validate and evaluate the performance of the model in predicting income levels. This validation process assesses how accurately the model can estimate income categories and provides a quantitative measure of its predictive capabilities. By analyzing the model's performance, HR professionals can gauge the reliability and effectiveness of the ordinal logistic regression in capturing and predicting income levels within their specific HR dataset. We can see that the ordinal logistic regression had the below variables in its final model.

Relaxed Lasso:

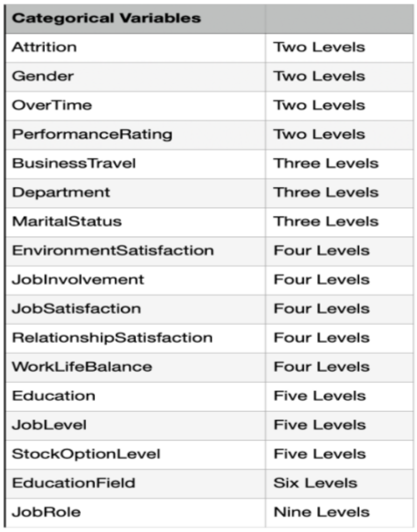
In response to the feedback, Jiten and I collaborated on implementing the relaxed lasso technique. Our primary goal was to develop a parsimonious model that effectively predicted attrition while maintaining strong performance. I took the lead in measuring the model's performance, ensuring a comprehensive evaluation. By incorporating the relaxed lasso method, we aimed to strike a balance between model simplicity and predictive accuracy. Together, Jiten and I carefully selected a subset of significant variables, optimizing the trade-off between complexity and performance. Our joint effort ensured a thorough examination of the model.

The project served as an excellent opportunity to integrate various techniques learned in class, including Multiple Correspondence Analysis and Ordinal Logistic Regression. It not only allowed for the application of familiar methods but also provided a platform to explore and implement new techniques. Overall, the project served as a valuable experience in combining and expanding our analytical repertoire.

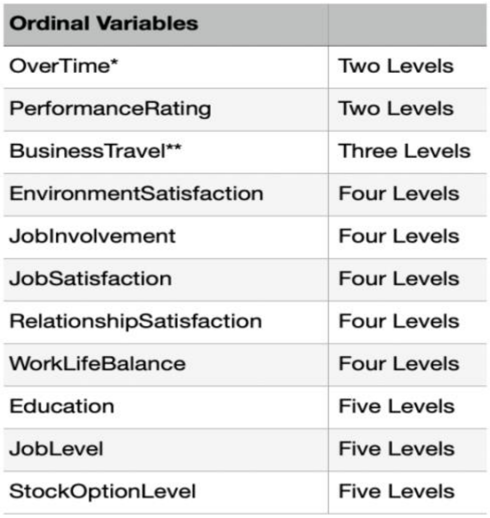
– Erik Pak

The primary objective of my analysis is to examine the similarities between seventeen categorical variables and understand their relationship to the response variable, which in this case is "Attrition." By applying MCA, my aim to uncover the underlying relationships and patterns among these variables and started with Correspondence Analysis (CA) leading into Multiple Correspondence Analysis (MCA).

Finally, exploring the possibility of deriving factors within each of the eleven ordinal variables by utilizing polychoric correlation matrix is significant enough to determine the factors in the ordinal variables.



Categorical Variables	
Attrition	Two Levels
Gender	Two Levels
OverTime	Two Levels
PerformanceRating	Two Levels
BusinessTravel	Three Levels
Department	Three Levels
MaritalStatus	Three Levels
EnvironmentSatisfaction	Four Levels
JobInvolvement	Four Levels
JobSatisfaction	Four Levels
RelationshipSatisfaction	Four Levels
WorkLifeBalance	Four Levels
Education	Five Levels
JobLevel	Five Levels
StockOptionLevel	Five Levels
EducationField	Six Levels
JobRole	Nine Levels

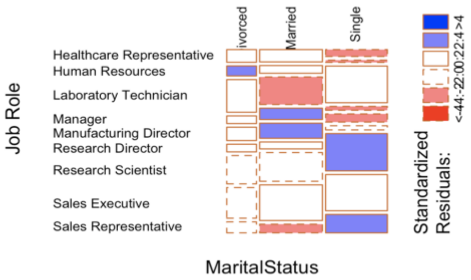


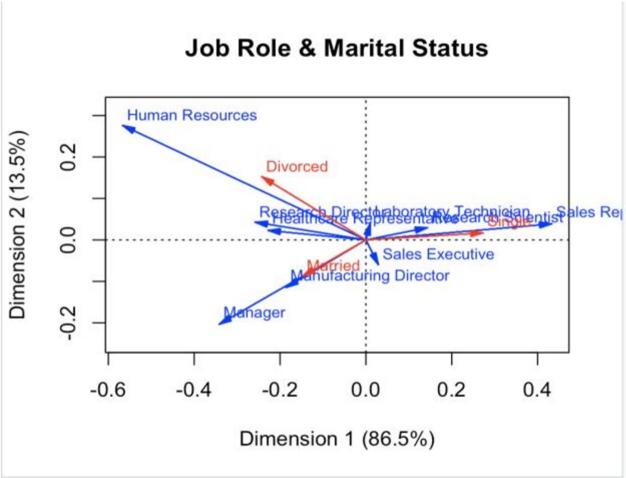
Ordinal Variables	
OverTime*	Two Levels
PerformanceRating	Two Levels
BusinessTravel**	Three Levels
EnvironmentSatisfaction	Four Levels
JobInvolvement	Four Levels
JobSatisfaction	Four Levels
RelationshipSatisfaction	Four Levels
WorkLifeBalance	Four Levels
Education	Five Levels
JobLevel	Five Levels
StockOptionLevel	Five Levels

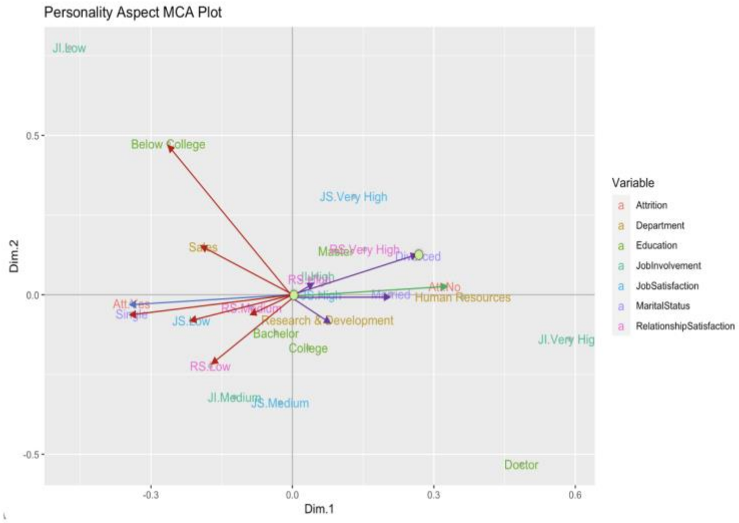
Correspondence Analysis (CA) is a valuable technique for analyzing the relationships between categorical variables. It allows me to visualize the associations between categories and identify patterns in a two-dimensional space. This helped me understand how the categorical variables are related and how they contribute to the response variable of interest, such as "Attrition" in my case.

Multiple Correspondence Analysis (MCA) is an extension of CA that can handle more than two categorical variables simultaneously. It allows me to explore the relationships and patterns among multiple categorical variables in a multidimensional space. MCA can provide a deeper understanding of the associations between variables and their connection to the response variable.

The Mosaic Plot on the left is a graphical representation that displays the joint distribution of categories across two categorical variables (Job Role & Marital Status). It provides a way to visualize the relationship and association between these variables in a rectangular grid which CA directly plots on the right. There is a direct relationship between the Mosaic plot and the CA biplot. For example, Human Resources & Divorce in the Mosaic plot shows a high positive standardized residual indicating that the observed count in a cell is higher than the expected count, which suggests a positive association between the categories of the two variables. That is reflected in the CA biplot stating divorce and Human Resources are closest to each other.







MCA biplot Indicate that individuals at the early stages of their careers are more likely to experience attrition. In contrast, mature individuals with a clearer understanding of their career path are less likely to leave their current job. This may be attributed to their knowledge, experience, and ability to make informed career decisions. Additionally, individuals with family responsibilities, such as children and a spouse, appear to have a lower likelihood of attrition due to their desire for stability and work-life balance. It is important to note that these findings are based on the observed patterns in the bi-plot, and further analysis is needed to validate and explore the complex factors influencing attrition in this dataset.

Based on the analysis conducted to explore the possibility of deriving factors from the eleven ordinal variables, it was determined that there was no significant linear relationship among these variables. This finding indicates that the variables do not exhibit a significant linear association.

As a result, it was concluded that each ordinal variable functions as an individual factor, and no underlying factors could be derived from the combination of these variables. Each variable is considered to represent a separate and distinct factor.

Learning Multiple Correspondence Analysis (MCA) was a challenging and exciting learning journey. MCA is a powerful and versatile technique for analyzing categorical data and exploring relationships between variables. I am excited about investigating and addressing the intricacies and nuances of Correspondence Analysis (CA). The more I delve into CA, the more questions arise, and my curiosity and desire to dig deeper into gaining a comprehensive understanding of the method and its potential applications is a tremendous feeling.

I felt like a child in a toy store and wished we had additional time to try further analysis. For example, using the coordinates or contributions of categories to dimensions obtained from MCA as input for clustering algorithms. Also, using multidimensional scaling (MDS) to analyze the dissimilarity or similarity between categorical variables by constructing a distance matrix based on the categorical data with those from MCA to gain insights into the relationships and similarities between the variables.